GEO: the Gene Expression Omnibus

A family of databases for gene expression related data

https://www.ncbi.nlm.nih.gov/geo/

National Center for Biotechnology Information • National Library of Medicine • National Institutes of Health • Department of Health and Human Services

Scope and Access

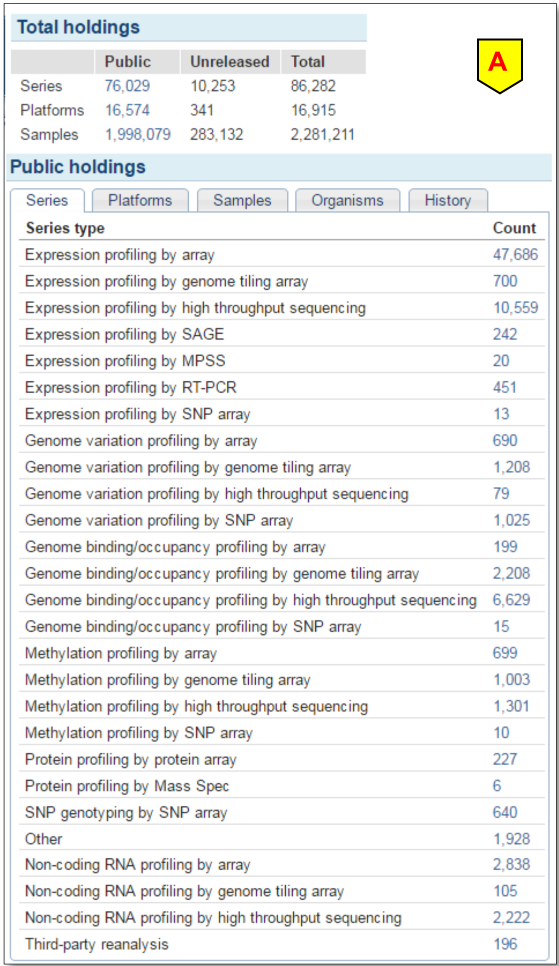
The Gene Expression Omnibus (GEO) is a public repository that archives and freely distributes microarray, next-generation sequencing, and other forms of high-throughput functional genomic data submitted by the scientific community. In addition to data storage, a collection of web-based interfaces, browse and search methods, and advanced analytical tools are available to help users find, analyze and download the data stored in GEO. The GEO homepage (https://www.ncbi.nlm.nih.gov/geo/) is the main gateway for searching and browsing high-throughput array data. GEO accepts many categories of high-throughput functional genomic data, including all array-based applications and some high-throughput sequencing data. Questions and issues related to submission should be sent to: geo@ncbi.nlm.nih.gov.

The GEO Databases, Data and Record Types

While GEO was originally established to host gene expression data, it has evolved to host other data types, including comparative genomic analyses, chromatin immunoprecipitation (ChIP) profiling that characterizes genome-protein interactions, non-coding RNA profiling, SNP genotyping and genome methylation status analyses. Thus, the data in GEO is measured in a wide variety of high-throughput assay methods (platforms) and measures a diverse set of sample types, as shown by the GEO Repository Browser (**A**, https://www.ncbi.nlm.nih.gov/geo/summary/) for the platforms.

Platform, Sample and Series records are submitted to the NCBI by researchers. A Platform record (**B**) is composed of a summary description of the array or sequencer and, for array-based Platforms, a data table defining the array template. A Sample record (**C**) is composed of a description of the biological material, the experimental protocols to which it was subjected, and a data table containing abundance measurements for each feature on the corresponding Platform table. A Series record (**D**) defines a set of related Samples considered to be part of a study, and describes the overall study aim and design.

DataSet records (**E**) are prepared by NCBI Curators with information extracted from the submitter-supplied records. A DataSet represents a summarized collection of consistently processed and experimentally related Sample records categorized according to experimental variables. Profiles (**F**) are derived from DataSets. A Profile consists of the expression measurements for an individual gene across all Samples in a DataSet. DataSets and Profiles are a means for transforming diverse styles of submitted data into a relatively standardized format. These curated database records form the basis of GEO's advanced data display and analysis tools.

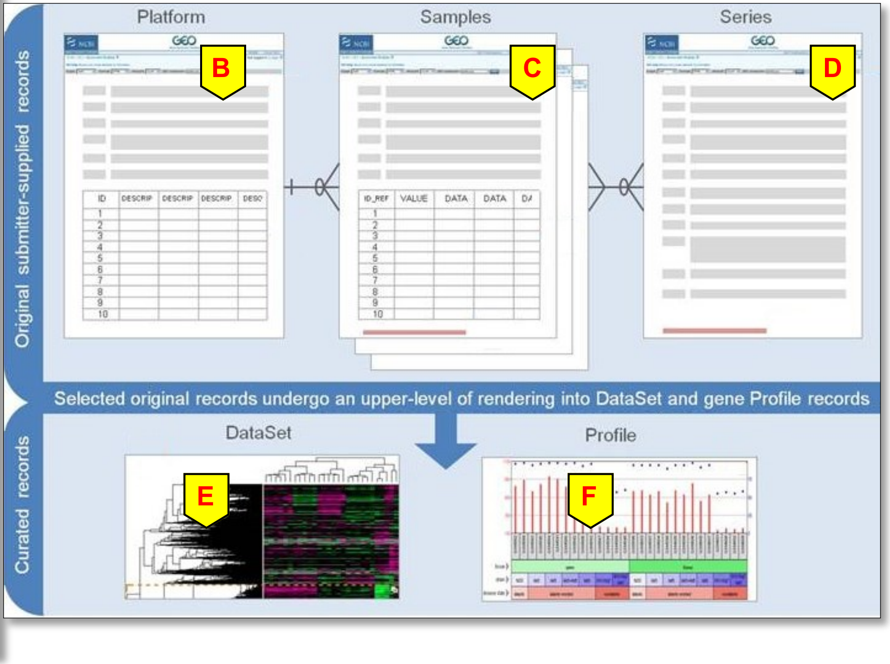


Total holdings

	Public	Unreleased	Total
Series	76,029	10,253	86,282
Platforms	16,574	341	16,915
Samples	1,998,079	283,132	2,281,211

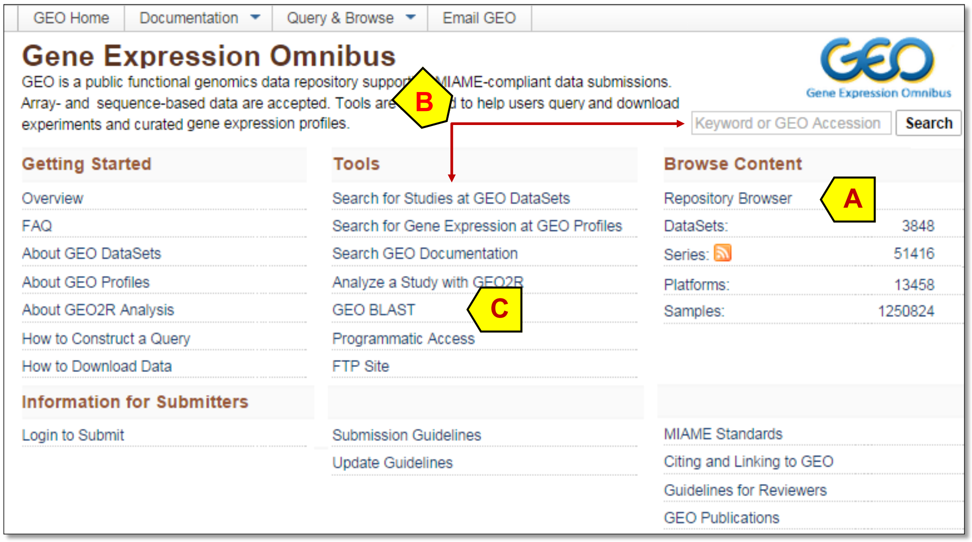
Public holdings

Series type	Count
Expression profiling by array	47,686
Expression profiling by genome tiling array	700
Expression profiling by high throughput sequencing	10,559
Expression profiling by SAGE	242
Expression profiling by MPSS	20
Expression profiling by RT-PCR	451
Expression profiling by SNP array	13
Genome variation profiling by array	690
Genome variation profiling by genome tiling array	1,208
Genome variation profiling by high throughput sequencing	79
Genome variation profiling by SNP array	1,025
Genome binding/occupancy profiling by array	199
Genome binding/occupancy profiling by genome tiling array	2,208
Genome binding/occupancy profiling by high throughput sequencing	6,629
Genome binding/occupancy profiling by SNP array	15
Methylation profiling by array	699
Methylation profiling by genome tiling array	1,003
Methylation profiling by high throughput sequencing	1,301
Methylation profiling by SNP array	10
Protein profiling by protein array	227
Protein profiling by Mass Spec	6
SNP genotyping by SNP array	640
Other	1,928
Non-coding RNA profiling by array	2,838
Non-coding RNA profiling by genome tiling array	105
Non-coding RNA profiling by high throughput sequencing	2,222
Third-party reanalysis	196


Original submitter-supplied records
Platform
Samples
Series
B
C
D
ID
DESCRIP
DESCRIP
DESCRIP
DESC
ID_REF
VALUE
DATA
DATA
DA
Selected original records undergo an upper-level of rendering into DataSet and gene Profile records
Curated records
DataSet
Profile
E
F

Browsing and Searching in GEO

In addition to being able to browse (A) GEO data in sortable, linked tables, GEO Profile and GEO DataSet databases can be searched using query boxes or from their individual homepages linked under the tools column (B). The Entrez search and retrieval system can be used to query for records corresponding to specific GEO accession numbers or for searching with terms for specific data attributes including keywords, authors, organisms, gene symbols, gene names, GenBank accession numbers, or Profiles flagged as being differentially expressed. Users can refine searches by simply typing in terms of interest or constructing very specific fielded queries. A full listing of all indexed fields, with examples and tips specifically for the GEO Databases, is provided at https://www.ncbi.nlm.nih.gov/geo/info/qqtutorial.html. Additionally, data for large collaborative projects, such as ENCODE and Roadmap Epigenomics, have dedicated browser pages at https://ww.ncbi.nlm.nih.gov/geo/info/ENCODE.html and https://www.ncbi.nlm.nih.gov/geo/roadmap/epigenomics/, respectively. Nucleotide or protein sequences can be used to search for genes in GEO Profiles through GEO BLAST (C). This page is also linked from the BLAST homepage (https://blast.ncbi.nlm.nih.gov/) under the “Specialized BLAST” section.



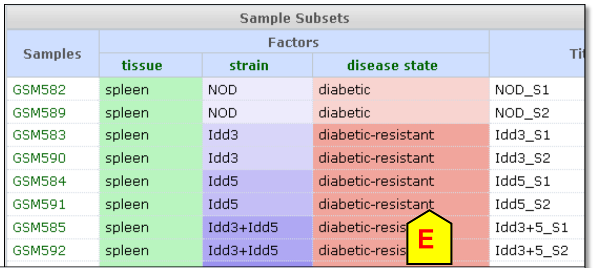
Gene Expression Omnibus

GEO is a public functional genomics data repository suppo IAME-compliant data submissions. Array- and sequence-based data are accepted. Tools are d to help users query and download experiments and curated gene expression profiles.

Getting Started	Tools	Browse Content	
Overview	Search for Studies at GEO DataSets	Repository Browser	
FAQ	Search for Gene Expression at GEO Profiles	DataSets:	3848
About GEO DataSets	Search GEO Documentation	Series:	51416
About GEO Profiles	Analyze a Study with GEO2R	Platforms:	13458
About GEO2R Analysis	GEO BLAST	Samples:	1250824
How to Construct a Query	Programmatic Access		
How to Download Data	FTP Site		
Information for Submitters			
Login to Submit	Submission Guidelines	MIAME Standards	
	Update Guidelines	Citing and Linking to GEO	
		Guidelines for Reviewers	
		GEO Publications	

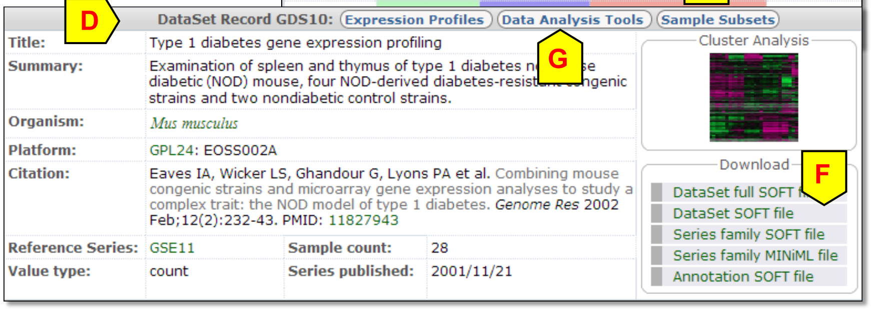
GEO DataSets Summary Pages

The GEO DataSets database stores and displays summary information about each curated experimental set as well as links to the underlying data, including the Platform, Series and other reference material such as organism and PubMed references The DataSet summary page for GDS10 is shown as an example (D). Samples included in the DataSet are linked from a sortable and fully linked table and which can be accessed with a click of the Sample Subsets button (E). Downloading the entire dataset can be accomplished quickly, in many formats (F). In addition, due to the curated nature of the data, Data Analytical Tools (G) are available for assessment of the data, including statistical comparison and cluster analysis of sample subsets.



Sample Subsets

Samples	Factors			Ti
	tissue	strain	disease state	
GSM582	spleen	NOD	diabetic	NOD_S1
GSM589	spleen	NOD	diabetic	NOD_S2
GSM583	spleen	Idd3	diabetic-resistant	Idd3_S1
GSM590	spleen	Idd3	diabetic-resistant	Idd3_S2
GSM584	spleen	Idd5	diabetic-resistant	Idd5_S1
GSM591	spleen	Idd5	diabetic-resistant	Idd5_S2
GSM585	spleen	Idd3+Idd5	diabetic-resis	Idd3+5_S1
GSM592	spleen	Idd3+Idd5	diabetic-resis	Idd3+5_S2

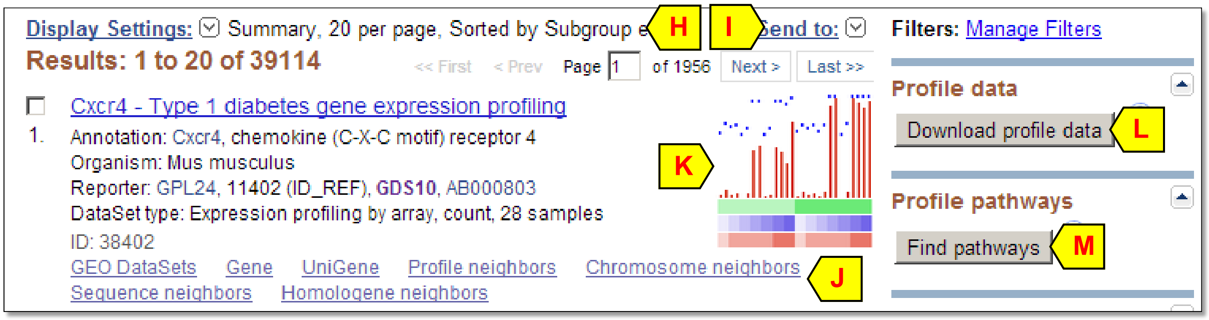


DataSet Record GDS10:

Title:	Type 1 diabetes gene expression profiling		
Summary:	Examination of spleen and thymus of type 1 diabetes n se diabetic (NOD) mouse, four NOD-derived diabetes-resist ngenic strains and two nondiabetic control strains.		
Organism:	*Mus musculus*		
Platform:	GPL24: EOSS002A		
Citation:	Eaves IA, Wicker LS, Ghandour G, Lyons PA et al. Combining mouse congenic strains and microarray gene expression analyses to study a complex trait: the NOD model of type 1 diabetes. *Genome Res* 2002 Feb;12(2):232-43. PMID: 11827943		
Reference Series:	GSE11	Sample count:	28
Value type:	count	Series published:	2001/11/21

GEO Profiles Summaries

The GEO Profile database stores gene expression profiles from curated DataSets. Retrieved records are displayed in summary format, which are sortable and downloadable through the Display Setting (H) and Send to (I) links, respectively. The summaries provide information about the particular gene and the Series or DataSet in which it was examined. They are fully linked within GEO and other NCBI databases (J). The Profile Neighbors link and Subgroup effect sort option provide access to other genes that exhibit similar expression behavior. Finally, the chart next to a profile displays the expression level of the gene for comparison of expression levels in all samples within a DataSet (K). Data points for retrieved profiles can be downloaded using the “Download profile data” button (L) and pathways involved can be retrieved from BioSystems using the “Find pathways” button (M).



Display Settings: Summary, 20 per page, Sorted by Subgroup e

Results: 1 to 20 of 39114

☐ Cxcr4 - Type 1 diabetes gene expression profiling

1. Annotation: Cxcr4, chemokine (C-X-C motif) receptor 4
Organism: Mus musculus
Reporter: GPL24, 11402 (ID_REF), GDS10, AB000803
DataSet type: Expression profiling by array, count, 28 samples
ID: 38402
GEO DataSets Gene UniGene Profile neighbors Chromosome neighbors Sequence neighbors Homologene neighbors

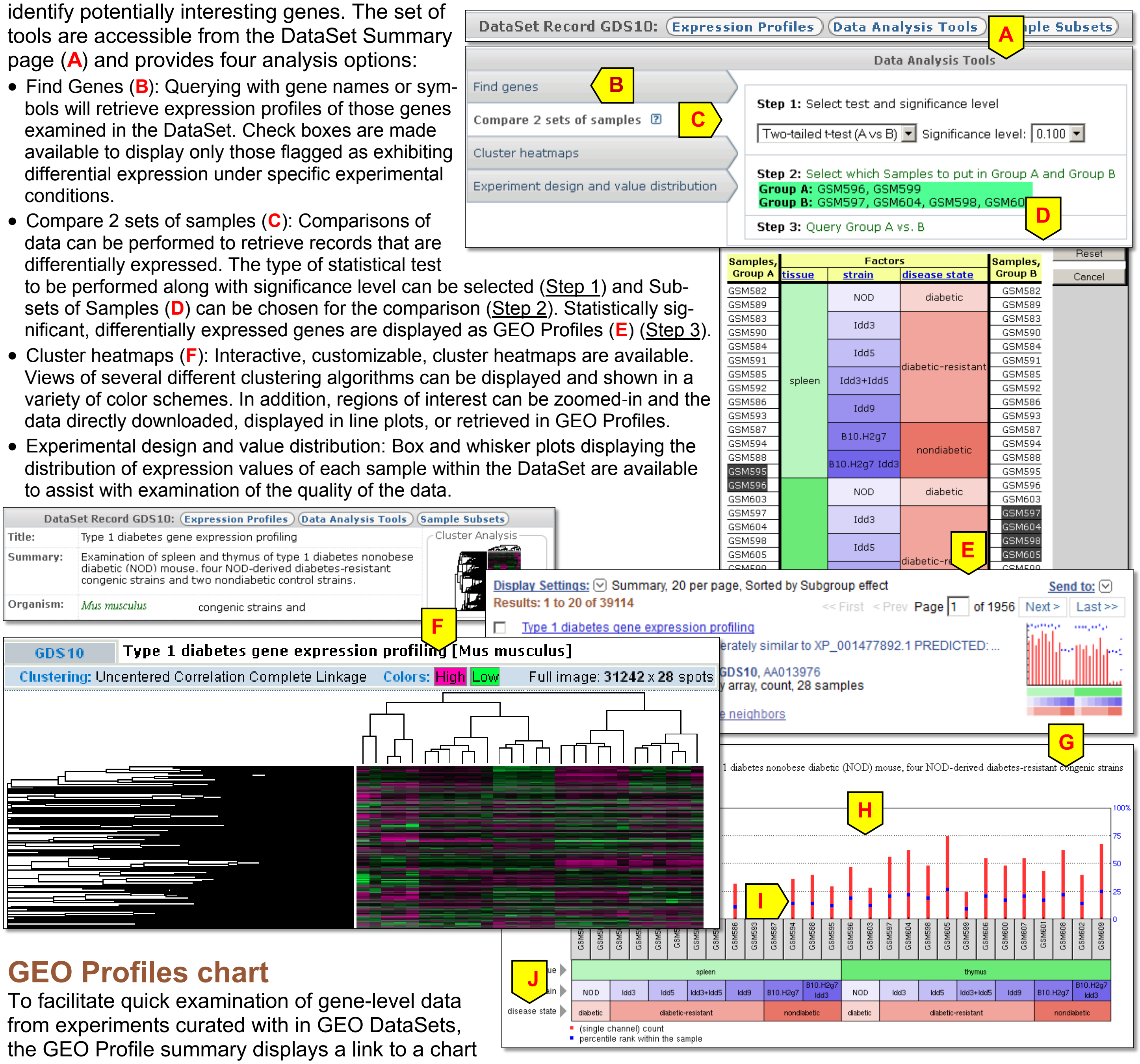
Analyzing with GEO2R

GEO2R is a simple web interface (https://www.ncbi.nlm.nih.gov/geo/geo2r/), which allows sophisticated R-based analysis of GEO data for the identification and visualization of differential gene expression, particularly for those data sets that were not curated by NCBI. Results are presented as a table of genes ordered by significance that can be visualized with GEO Profile graphics. A YouTube video tutorial is available at https://www.youtube.com/watch?v=EUPmGWS8ik0

Customizable GEO DataSet data analysis tools

GEO DataSet records have integrated Data Analysis Tools to facilitate examination and interrogation of the data to help identify potentially interesting genes. The set of tools are accessible from the DataSet Summary page (**A**) and provides four analysis options:

- Find Genes (**B**): Querying with gene names or symbols will retrieve expression profiles of those genes examined in the DataSet. Check boxes are made available to display only those flagged as exhibiting differential expression under specific experimental conditions.
- Compare 2 sets of samples (**C**): Comparisons of data can be performed to retrieve records that are differentially expressed. The type of statistical test to be performed along with significance level can be selected (Step 1) and Subsets of Samples (**D**) can be chosen for the comparison (Step 2). Statistically significant, differentially expressed genes are displayed as GEO Profiles (**E**) (Step 3).
- Cluster heatmaps (**F**): Interactive, customizable, cluster heatmaps are available. Views of several different clustering algorithms can be displayed and shown in a variety of color schemes. In addition, regions of interest can be zoomed-in and the data directly downloaded, displayed in line plots, or retrieved in GEO Profiles.
- Experimental design and value distribution: Box and whisker plots displaying the distribution of expression values of each sample within the DataSet are available to assist with examination of the quality of the data.


DataSet Record GDS10: Expression Profiles | Data Analysis Tools | ...ple Subsets
A
Data Analysis Tools
Find genes
B
Compare 2 sets of samples
C
Cluster heatmaps
Experiment design and value distribution
D
Reset
Cancel
E
F
G
H
I
J

Step 1: Select test and significance level

Two-tailed t-test (A vs B) ▾ Significance level: 0.100 ▾

Step 2: Select which Samples to put in Group A and Group B

Group A: GSM596, GSM599

Group B: GSM597, GSM604, GSM598, GSM60

Step 3: Query Group A vs. B

Samples, Group A	tissue	strain	disease state	Samples, Group B
GSM582	spleen	NOD	diabetic	GSM582
GSM589				GSM589
GSM583		Idd3	diabetic-resistant	GSM583
GSM590				GSM590
GSM584		Idd5		GSM584
GSM591				GSM591
GSM585		Idd3+Idd5		GSM585
GSM592				GSM592
GSM586		Idd9		GSM586
GSM593				GSM593
GSM587		B10.H2g7	nondiabetic	GSM587
GSM594				GSM594
GSM588		B10.H2g7 Idd3		GSM588
GSM595				GSM595
GSM596		NOD	diabetic	GSM596
GSM603				GSM603
GSM597		Idd3	diabetic-re	GSM597
GSM604				GSM604
GSM598		Idd5		GSM598
GSM605				GSM605
GSM599				GSM599

DataSet Record GDS10: Expression Profiles | Data Analysis Tools | Sample Subsets

Title: Type 1 diabetes gene expression profiling

Summary: Examination of spleen and thymus of type 1 diabetes nonobese diabetic (NOD) mouse. four NOD-derived diabetes-resistant congenic strains and two nondiabetic control strains.

Organism: *Mus musculus* congenic strains and

Cluster Analysis

GDS10 Type 1 diabetes gene expression profiling [Mus musculus]

Clustering: Uncentered Correlation Complete Linkage Colors: High Low Full image: 31242 x 28 spots

Display Settings: Summary, 20 per page, Sorted by Subgroup effect Send to:

Results: 1 to 20 of 39114

<< First < Prev Page 1 of 1956 Next > Last >>

□ Type 1 diabetes gene expression profiling

...erately similar to XP_001477892.1 PREDICTED: ...

GDS10, AA013976

...y array, count, 28 samples

...e neighbors

...1 diabetes nonobese diabetic (NOD) mouse, four NOD-derived diabetes-resistant congenic strains

GEO Profiles chart

To facilitate quick examination of gene-level data from experiments curated with in GEO DataSets, the GEO Profile summary displays a link to a chart analysis of gene expression levels. Clicking on the Profiles thumbnail (**G**) displays a larger version of the chart (**H**). Full profile details are shown, including descriptions of and links to the DataSet and Samples, as well as the expression profile for the individual gene grouped by experimental subset. The red bars indicate the level of abundance of an individual transcript across the Samples that make up a DataSet, while blue squares indicate rank order for that gene's expression with respect to the other genes within the Sample (**I**). The colored X-axis bars (**J**) distinguish named experimental subsets of Samples within the DataSet (type, e.g., 'tissue', description, e.g., 'spleen', and disease state, e.g., 'diabetic'). Located under the chart is a button to display data values, as well as a link to a help document with thorough explanation of this view.

GEO Submission

GEO accepts many categories of high-throughput functional genomic data, including all array-based applications and some high-throughput sequencing data. The GEO Curation Team is available to assist submitters in depositing their data. Send an E-mail to: geo@ncbi.nlm.nih.gov with a brief description of the type of data you are trying to submit and an explanation of any problems with or questions about the submission procedures. Three different methods for submission available are summarized in the table below.

Submission Method	Format	Key points
GEOarchive	Spreadsheets (e.g., Excel)	Recommended method for most submissions. Quick description of the submitted study or studies using Excel spreadsheet templates.
SOFT	Plain Text	Good option if data and metadata are already in a database which can generate and export data in SOFT plain text format.
MINiML	XML	Good option if data and metadata are already in a database which can generate and export data in MINiML XML format.

All deposit methods described here support the submission of many data types, including: Gene expression, SNP arrays, SAGE, ChIP-chip, Protein arrays, ArrayCGH (Comparative Genomic Hybridization) and high throughput quantitative sequence data. In addition, submissions may remain private until a manuscript describing the data is published. However, as manuscript reviewers often need access to the data for evaluation purposes, GEO has developed a system for creation of a Reviewer URL, providing confidential and anonymous access, which authors can supply in their manuscript submissions. More on GEO submission is available at https://www.ncbi.nlm.nih.gov/geo/info/submission.html.

Downloading Bulk Data Files and Programmatic Access

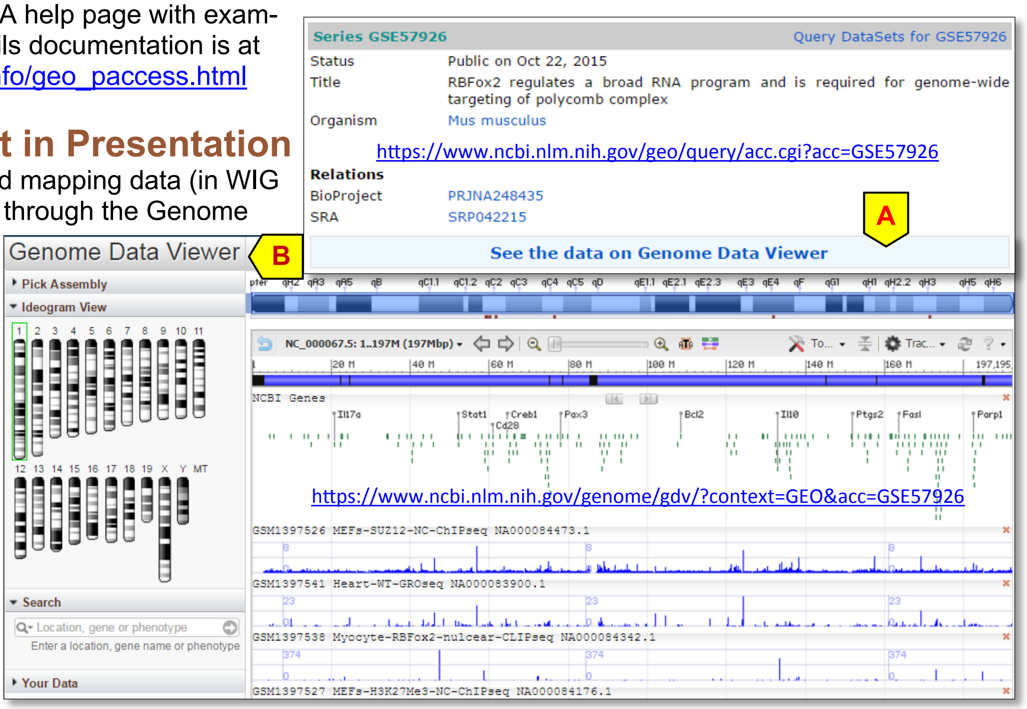
Data from GEO DataSets and Profiles databases, including all Platform, Sample, Series and DataSet records as well as raw data, are available for bulk download through FTP at ftp://ftp.ncbi.nih.gov/pub/geo/. In addition, metadata from these GEO-specific databases can be accessed programmatically using a suite of programs collectively referred to as the Entrez Programming Utilities (EUtils). A help page with examples and a link to the complete EUtils documentation is at https://www.ncbi.nlm.nih.gov/geo/info/geo_paccess.html

A Notable Enhancement in Presentation

GEO now makes submitter provided mapping data (in WIG and BEDgraph formats) accessible through the Genome Data Viewer. Datasets with this enabled are readily retrievable using a fielded query term **'has track"[prop]**. Clicking the " See the data on Genome Data Viwer" button (**A**) of an entry with this enabled brings out this display for that dataset (**B**).

GEO Documents and References

The collection GEO help documents on data submission, searching, linking, and data analysis are linked from the GEO homepage (https://www.ncbi.nlm.nih.gov/geo) under the "Getting Started" column.



Series GSE57926 Query DataSets for GSE57926

Status Public on Oct 22, 2015
Title RBFox2 regulates a broad RNA program and is required for genome-wide targeting of polycomb complex
Organism Mus musculus

https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE57926

Relations
BioProject PRJNA248435
SRA SRP042215

See the data on Genome Data Viewer

Genome Data Viewer

NC_000067.5: 1..197M (197Mbp)

https://www.ncbi.nlm.nih.gov/genome/gdv/?context=GEO&acc=GSE57926

GSM1397526 MEFs-SUZ12-NC-ChIPseq NA000084473.1
GSM1397541 Heart-WT-GROseq NA000083900.1
GSM1397538 Myocyte-RBFox2-nulcear-CLIPseq NA000084342.1
GSM1397527 MEFs-H3K27Me3-NC-ChIPseq NA000084176.1

A
B
Pick Assembly
Ideogram View
1 2 3 4 5 6 7 8 9 10 11
12 13 14 15 16 17 18 19 X Y MT
Search
Location, gene or phenotype
Enter a location, gene name or phenotype
Your Data
pter qA2 qA3 qA5 qB qC1.1 qC1.2 qC2 qC3 qC4 qC5 qD qE1.1 qE2.1 qE2.3 qE3 qE4 qF qG1 qH1 qH2.2 qH3 qH5 qH6
To...
Trac...
1 20 M 40 M 60 M 80 M 100 M 120 M 140 M 160 M 197,195
NCBI Genes
Il17a
Stat1
Cd28
Creb1
Pax3
Bcl2
Il10
Ptgs2
Fasl
Parp1
8
0
23
374

GEO Publications:

1. NCBI GEO: archive for functional genomics data sets - 10 years on. Barrett T, et.al. Nucleic Acids Res. 2011 Jan;39 (Database issue):D1005-10. Epub 2010 Nov 21. https://www.ncbi.nlm.nih.gov/pubmed/21097893
2. NCBI GEO standards and services for microarray data. Edgar R, Barrett T. Nat Biotechnol. 2006 Dec;24(12):1471-2. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2270403/.
3. Gene expression omnibus: microarray data storage, submission, retrieval, and analysis. Barrett T, Edgar R. Methods Enzymol. 2006;411:352-69. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1619900.

See more at https://wwww.ncbi.nlm.nih.gov/geo/info/citations.html.